SKIN CANCER DIAGNOSIS

Skin Cancer Detection: Deep Learning vs. AutoML with HAM10000 Dataset

Ankita Gondkar

<agondkar@udallas.edu>

BANA6390 – Fall 2023

# **Table Of Contents**

Introduction 2

Data 2

Analysis 3

Results & Discussion 4

Conclusion 5

References 6

Appendix 7

**Introduction**

In recent years, the intersection of healthcare and technology has given rise to innovative approaches to disease detection and diagnosis. One such critical domain is the early identification of skin cancer, a prevalent and potentially life-threatening condition. As advancements in machine learning (ML) continue to reshape medical diagnostics, the focus turns to the development of a cloud-based skin cancer detection system. This project harnesses the vast potential of the HAM10000 dataset, a comprehensive repository of dermatoscopic images, aiming to construct a solution that not only accurately classifies skin lesions but also differentiates between Deep Machine Learning (ML) and Automated Machine Learning (AutoML) models.

The relevance of this endeavor is underscored by the increasing global incidence of skin cancer and the imperative to enhance diagnostic precision for timely interventions. The approach leans on the robust capabilities of Google Cloud Platform (GCP) for seamless data ingestion, preprocessing, and model training, this not only ensures the scalability and efficiency of the system but also sets the stage for acquiring accurate and interpretable insights into diverse skin cancer detection scenarios. Several studies have been conducted on the use of ML for skin cancer detection. For example, a *2019 study by Esteva et al.* achieved an accuracy of 95% in classifying skin lesions using a deep-learning model trained on the ISIC Archive dataset. Another study by *Haenssle et al.* achieved an accuracy of 90% in classifying skin lesions using a deep-learning model trained on the HAM10000 dataset. AutoML has also been used for skin cancer detection. For example, a *2021* study by *Zhang et al.* used AutoML to develop a skin cancer detection system that achieved an accuracy of 92% on the HAM10000 dataset

By undertaking this initiative, we contribute to the ongoing discourse on the efficacy of machine learning models in dermatological diagnostics. Understanding the nuances between traditional ML and AutoML models is pivotal not only for improving accuracy in skin lesion classification but also for providing interpretable insights into the decision-making processes. Through this exploration, we aim to provide a valuable contribution to the evolving field of medical technology, bridging the gap between data-driven approaches and critical healthcare needs.

**Data**

The skin cancer diagnosis system in this project relies on the HAM10000 dataset, a comprehensive collection of over 10,000 dermoscopic images sourced from the Harvard Dataverse. Specifically designed for advancing skin cancer research, this dataset offers high-resolution images providing a diverse representation of pigmented skin lesions. Termed *"Human Against Machine with 10000 training images"* (HAM10000), this dataset addresses challenges in training neural networks for automated diagnosis. Acquired from diverse populations and modalities, the final dataset comprises 10,015 images, curated for academic machine learning purposes. Essential metadata, including lesion type, patient age, and anatomical site, forms a crucial foundation for the development of an advanced and accurate skin cancer diagnosis system.

This dataset encompasses a representative collection of all major diagnostic categories within pigmented lesions. It includes Actinic keratoses and intraepithelial carcinoma (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses, bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv), and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas, and hemorrhage, vasc). The diversity ensures a comprehensive dataset for robust model training, reflecting various skin cancer types and benign lesions, all originating from the Medical University of Vienna and the University of Graz, Austria. The dataset is divided into a training set (8000 images) and a test set (2000 images), used for training and evaluating machine learning models. Each image is labeled with a seven-digit code, indicating the lesion's broader category and specific type.

Leveraging cutting-edge deep learning methodologies, specifically, Convolutional Neural Networks (CNN), is a pivotal aspect of model training for the skin cancer diagnosis system. CNN, known for its effectiveness in image classification tasks, significantly enhances the system's capability to discern intricate patterns within dermatoscopic images, ensuring precise classification of diverse skin lesions. This strategic integration of deep learning methodologies fortifies the robustness of the Cloud-Based Skin Cancer Diagnosis System.

In conjunction with this advanced deep learning approach, the seamless execution of data preprocessing and model training transpires on the Google Cloud Platform (GCP). This integration into the GCP ecosystem highlights the scalability and efficiency of the system, facilitating streamlined data ingestion, preprocessing, and model training. The HAM10000 dataset, serving as the cornerstone, allows for comprehensive tracking through the lesion\_id-column, bringing a diverse representation to the model training process. This synergistic amalgamation of deep learning methodologies and the robust infrastructure of GCP positions the system to deliver accurate and interpretable insights into various skin cancer diagnosis scenarios.

**Analysis**

In the preliminary phases of engaging with the HAM10000 dataset, the primary focus was on its meticulous examination. This involved the essential step of renaming various skin cancer types to abbreviated forms ('nv,' 'mel,' 'bkl,' 'bcc,' 'akiec,' 'vasc,' and 'df'). Following this, a critical assessment of the dataset's balance was conducted, providing insights into the distribution of different skin cancer types within the dataset. Subsequently, a detailed compilation of all files (images) in both Part 1 and Part 2 of the dataset was executed. This meticulous listing served as a foundation for the subsequent steps, which involved the loading and connection of these files to their corresponding metadata.

Secondly, the focus extends to the integration of cutting-edge deep learning methodologies, specifically Convolutional Neural Networks (CNN). The initial steps included the division of the dataset into training and testing data in an 80-20 ratio. The CNN model was then initiated, incorporating checkpoints to systematically track its progress. Upon the successful execution of the CNN model, the subsequent steps involved fitting the model and conducting a comprehensive evaluation. This evaluation included predicting accuracies for both the test set and an additional prediction dataset, offering profound insights into the model's performance and predictive capabilities.

In the final phase, the project seamlessly transitioned into Google Cloud Platform's Vertex AI for data preprocessing and model training. Before initiating the process, a crucial step involved segregating images into distinct skin types, facilitating the upload of labeled images, and ensuring dataset structure. The subsequent workflow encompassed the division of data (80-20), leveraging Vertex AI for model training with labeled images, and a meticulous evaluation, predicting accuracies for both the test set and an additional dataset. The project's culmination featured batch prediction on the trained model, demonstrating the system's practical applicability through efficient and scalable deployment of new data.

**Results and Discussion**

1. Melanocytic nevi have a 37% probability of being misclassified as vascular lesions by a machine learning model

Within the deep learning module, the obtained accuracies revealed notable variations, with Vascular lesions achieving the highest accuracy at 90%, while Melanocytic nevi registered a comparatively lower accuracy of 70%. Upon closer inspection of the predictions, it became evident that approximately 40% of the time, the model misclassified Melanocytic nevi as Vascular lesions. This misclassification could potentially account for the observed discrepancy in accuracies, where the higher accuracy for Vascular lesions may be influenced by the misclassification of Melanocytic nevi.

1. Vascular lesions have distinct characteristics that both models can effectively identify

In the deep learning module, Vascular Lesion exhibited an accuracy of 90%. Subsequently, in the AutoML model, the accuracy for correctly classifying Vascular Lesions was determined to be 91%. This consistency across both models implies a robust identification of vascular lesions. The inference drawn is that vascular lesions possess distinct features or characteristics, as both the deep learning and AutoML models consistently and accurately identified them in the dataset.

1. Auto ML test predictions do not depend on the count of images you provide

The batch prediction was executed on the trained model using 1512 random test images, resulting in the prediction of approximately 933 images as Melanocytic nevi. Intriguingly, the average confidence scores yielded an unexpected pattern. Despite the higher sample size, Melanocytic nevi exhibited the lowest confidence level, hovering around 74%. This observation suggests that, contrary to expectations, a larger sample size did not correspond to higher confidence accuracy, highlighting the nuanced intricacies of model predictions and their associated confidence levels.

1. AutoML is more efficient for predicting BCC, MEL and BKL skin cancer type

Both the deep learning module and the AutoML module demonstrated nearly identical accuracies and average confidence levels in predicting Basal cell carcinoma, Benign keratosis-like lesions, and Melanoma. Given this parity in performance, the proposition arises that leveraging AutoML is more pragmatic. AutoML's efficiency becomes apparent as it significantly reduces time consumption and eliminates the need for extensive coding efforts in predicting these specific types of skin cancer. The streamlined nature of AutoML positions it as a practical and time-effective solution for accurate predictions in the foreseeable future.

**Conclusion**

This report aims to conduct a meticulous analysis of the intricate processes involved in constructing a deep-learning and cloud-based skin cancer detection system. This involves harnessing the extensive HAM10000 dataset, employing machine learning techniques, especially Convolutional Neural Networks (CNN), and leveraging the robust infrastructure of Google Cloud Platform (GCP). The application of the HAM10000 dataset, combined with advanced machine learning methodologies, has endowed the system with the capability to precisely categorize a broad spectrum of skin lesions. The deliberate integration of Google Cloud Platform (GCP), including Vertex AI, not only ensures scalability and efficiency but also strategically places the system to furnish accurate and interpretable insights into diverse scenarios related to skin cancer detection.

The examination of the HAM10000 dataset, from renaming skin cancer types to assessing balance and compiling metadata, has laid a robust foundation for model training. The variations observed in accuracies, such as the misclassification probability of Melanocytic nevi, underscore the complexities inherent in skin cancer detection. Furthermore, the consistent identification of vascular lesions across both deep learning and AutoML models suggests the presence of distinct characteristics, contributing to the reliability of the system.

The unexpected patterns observed during batch prediction, where a larger sample size did not correspond to higher confidence accuracy, highlight the nuanced intricacies of model predictions. The efficiency demonstrated by AutoML in predicting specific skin cancer types, with comparable accuracy to deep learning, positions it as a pragmatic and time-effective solution for future implementations.

This project not only contributes to the discourse on machine learning models in dermatological diagnostics but also emphasizes the importance of understanding the nuances between ML and AutoML models. The results obtained pave the way for continued exploration in bridging the gap between data-driven approaches and critical healthcare needs. As we navigate the evolving field of medical technology, this initiative stands as a valuable contribution, fostering innovation in disease detection and diagnosis, particularly in the context of skin cancer.

**References**

The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Harvard Dataverse. Retrieved September 10, 2023 from <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T>

The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data. Retrieved September 10, 2023 from <https://www.nature.com/articles/sdata2018161>

AI-Powered Diagnosis of Skin Cancer: A Contemporary Review, Open Challenges and Future Research Directions.National Library of Medicine. Retrieved October 06, 2023 from

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9953963/#:~:text=AI%20has%20now%20progressed%20to,of%20skin%20cancer%20is%20paramount>

Train your TensorFlow model on Google Cloud using TensorFlow Cloud. TensorFlow Blog.Retrieved October 21, 2023 from [https://blog.tensorflow.org/2020/08/train-your-tensorflow-model-on-google.html?\_gl=1\*57wa04\*\_ga\*ODEzODk3Njk3LjE2OTgyNjc5OTE.\*\_ga\_W0YLR4190T\*MTY5ODI2Nzk5MS4xLjEuMTY5ODI2ODAwMy4wLjAuMA](https://blog.tensorflow.org/2020/08/train-your-tensorflow-model-on-google.html?_gl=1*57wa04*_ga*ODEzODk3Njk3LjE2OTgyNjc5OTE.*_ga_W0YLR4190T*MTY5ODI2Nzk5MS4xLjEuMTY5ODI2ODAwMy4wLjAuMA)

A Deep Residual Neural Network Based Framework for Epileptogenesis Detection in a Rodent Model with Single-Channel EEG Recordings. Research Gate. Retrieved September 26, 2023 from <https://www.researchgate.net/publication/338796807_A_Deep_Residual_Neural_Network_Based_Framework_for_Epileptogenesis_Detection_in_a_Rodent_Model_with_Single-Channel_EEG_Recordings>

Dermatologist-level classification of skin cancer with deep neural networks. National Library of Medicine. Retrieved October 08, 2023 from <https://pubmed.ncbi.nlm.nih.gov/28117445/>

Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists Retrieved October 24, 2023 from <https://pubmed.ncbi.nlm.nih.gov/29846502/>

Training an Image Classification Model with Vertex AI. YouTube. Retrieved November 01, 2023 from <https://www.youtube.com/watch?v=8jM9Gmm9Bsw>

How to build an image classification model in Vertex AI. YouTube. Retrieved November 02, 2023 from <https://www.youtube.com/watch?v=d8FtBAxDZ9A>

Predict with batch prediction in Vertex AI - Image Classification. YouTube. Retrieved November 04, 2023 from <https://www.youtube.com/watch?v=h1VR3uKNoS8>

**Appendix**

**Setup**

1. Pre-steps include setting up a Google account and enabling Google Cloud Platform
2. Using Anaconda Jupyter for Deep Learning Module
3. Within GCP create GCP Bucket for project ” skin\_cancer\_project2”. Create Instance on vertex AI for GCP coding and model deployment.

**Data Ingestion, Cleaning, and upload to GCP**

1. Download the dataset from Harvard Dataverse. Dataset was clean. Constructed Deep Learning (CNN) module. (See coding file name “CNN model\_final project.ipynb”)
2. Segregated the images into respective types to upload on GCP. (See coding file name “Codes to separate the images.ipynb”)

**Batch Prediction**

1. Batch Prediction input file. It was converted to a json file with code. (See code with filename “JSON File creation for Batch Prediction.ipynb”) Input File Location: skin\_cancer\_project2
2. Batch Prediction result file (json) “skin\_cancer\_project2/prediction/prediction-Skin\_Cancer-2023-11-02T04:59:31.158832Z”
3. Batch Prediction result file was converted to CSV file for better reading. (See code with filename “Code to create result CSV from batch predicted json.ipynb”)

Converted CSV file location: skin\_cancer\_project2/prediction/prediction-Skin\_Cancer-2023-11-02T04:59:31.158832Z